INSIDER'S GUIDE TO Data Protection: What It Is, and Finding the Right Provider

Deciding what to protect, and how, requires careful planning. We'll help you sort through the options. **By Trevor Pott**

If your data doesn't exist in two places, then it doesn't exist. This is an inescapable fact of information technology. With the rise of ransomware and other modern threats, it's even fair to say that if your data doesn't exist in multiple versioned copies located in two places that have separate administration planes, then your data does not exist. Welcome to the brave new world where your data may never be safe again.

Backups, disaster recovery (DR), continuous data protection (CDP), high availability (HA), fault tolerance and automated failover/failback are technical terms relevant to various parts of the disaster preparedness and recovery conversation. At one time, each of these had a specific meaning. Over time, common usage, changes in technology and the use as marketing buzzwords has blurred the lines between them.

In today's world only two terms really matter: data protection and workload protection. Achieving data protection requires the use of multiple techniques and is a must for everyone. Workload protection, on the other hand, if frequently a nice-to-have, can be either simple or complex, depending on needs.

Given the complexity of requirements, implementations and the constant evolution of this space, organizations are increasingly turning to managed or hosted services. Understanding why—and picking the right one—requires a brief primer in the relevant concepts.

SHUTTERSTOCK

Workload Protection

Workload protection is the easier of the two topics to discuss. There really aren't that many ways to go about it, or that many reasons an organization might want it.

At its heart, workload protection is about making sure that IT workloads that have failed for one reason or another continue to be available for use. The three main methods of workload protection are HA, fault tolerance and failover/failback.

HA and fault tolerance rely on multiple, different physical servers having access to the same storage. In the case of HA, if one physical server fails, then all the workloads running on that server will cease to operate. Some mechanism (which varies depending on the vendor) will detect the failure of these workloads and direct a different physical server to restart those workloads.

HA results in downtime, however brief, between detection of a server failure and the workloads being brought online again. Usually it's roughly as long as it takes to boot up a server.

Fault tolerance has a slightly different approach. With fault tolerance, the same workload is run on multiple computers simultaneously, with each computer being kept in lock-step with the others. If one server dies, the others take over and service isn't interrupted.

HA can lead to minor data loss. If there was data to be written to the disk that had not yet been committed to disk, then that data would be lost when the primary server fails. For many workloads, this doesn't matter. For others, it's critical.

Let's say that the workload in question was a Web site. If the site was mostly for information purposes, then the only likely data loss is a few seconds of log information about who's visiting the site and what they're accessing.

If, however, that site was actively collecting information, HA could be a problematic workload protection scheme. Imagine that the site in question was an electronic voting Web site that failed on election night. Votes may have been cast by visitors to the site, but the database behind it might not have committed all the information to disk yet. Here, fault tolerance is called for.

Failover and failback add a layer of complexity to the calculations. They typically refer to the ability to recover from the loss of the entire datacenter; they're related to failure recovery across vast geographic distances. Here, the laws of physics get in the way.

> **HA and fault tolerance are usually schemes that exist within a single, physical premises.**

HA and fault tolerance are usually schemes that exist within a single, physical premises. They rely on shared storage between the physical servers, and only a handful of storage vendors can even provide real-time replication at distances of 60 miles over fiber optic connections.

The more geographically distant the locations to be kept in sync, the more physics gets in the way. The speed of light is absolute and eventually the latency it creates would render workloads unusable.

That means failover and failback can have data loss, or not, depending on circumstances. In the case of a controlled failover, some technologies allow workloads to synchronize by temporarily halting the primary workload and then cutting over. Interruptions to the workload in this case can be as short as 100ms.

Where failover occurs because of an unexpected loss of the primary site (usually due to someone going through the primary site's Internet connectivity with a backhoe), workloads are likely to have to be restarted and will likely experience data loss.

CDP-Based Data Protection

Data protection comes in different flavors. Backup, DR and CDP are frequently discussed topics, but even these only barely scratch the surface. There are so many terms and technologies that fit under the umbrella of data protection that it's pointless to try to discuss them all. What matters instead is what you're trying to protect your data against.

Different products and services are used to protect against different failure domains. RAID and replication can help protect against the failure of physical hardware. Offline data copies (such as backing up to tape) can help protect against deletion, while off-site copies of data (usually grouped under the banner of DR) help protect against the loss of a datacenter to fire, natural disaster or the guy with the backhoe.

CDP is an important concept in data protection. At its most basic, it means ensuring that two copies of the data exist on different devices—sometimes in different datacenters—with a a recovery point objective (RPO) as close to zero as possible. An RPO is, in essence, "How much data can we afford to lose?"

This is also known as replication, and it means every single change is sent from the primary storage device to the secondary. CDP is all about being able to fail over workloads with minimal data loss. Unfortunately, CDP is typically expensive, and doesn't cover all data protection requirements.

With most CDP setups, if something is deleted on the primary storage, it's deleted on the secondary storage instantaneously, as well. As such, CDP usually provides zero protection against accidental deletion and other forms of "Oopsie McFumblefingers" human errors.

Corruption is another problem for CDP. Application crashes, ransomware and more can render data unusable, requiring a reversion to a previous version. If the CDP in use is focused merely on replication of data, corruption will spread from the primary storage to the secondary as quickly as deletions do.

Over the years, some CDP implementations have utilized a complete transaction history that allows rolling back individual files, objects and block storage devices on a write-by-write basis to any arbitrary point in time. This is fantastically hard to do well, requires exceptionally high-end storage gear and the logging tends to be so write-intensive that it will wear out SSDs on the destination site in short order. Very few CDP implementations use this today.

More frequently, CDP implementations allow you to set RPOs for individual file shares, object stores and block storage devices. These use replication to keep blocks in sync between storage devices, and then take regular snapshots to provide the desired RPOs.

Of course, CDP isn't magic. As discussed before, there are real-world limits to replication. The farther apart the source and destination, the longer the gap between a storage transaction occurring on the source and being recorded on the destination. The network bandwidth between the two storage devices also matters. This leaves room for non-CDP data protection.

> CDP usually provides zero protection against accidental deletion and other forms of "Oopsie McFumblefingers" human errors.

Non-CDP-Based Data Protection

The biggest downside to CDP-based data protection mechanisms is the cost of sending each and every change down the wire. Bandwidth isn't cheap, even inside a datacenter. It gets significantly less so when talking about data protection between physical sites.

Non-CDP-based data protection takes a different approach. Instead of streaming a copy of the data to the secondary storage device and then snapshotting at the destination, snapshots are taken on the primary storage device; only those snapshots are sent to the secondary device.

The result is usually dramatically lower bandwidth usage. More often than not, the information that changes between snapshots involves changing the same blocks of data several times, so only the final result of all those changes is sent at the time of the snapshot, instead of all the incremental changes in between snapshots.

CDP sends changes as they happen; this can get messy if there's contention for the bandwidth used. (For example, because you're attempting data protection over an Internet link.) Non-CDP data protection also allows scheduling of data synchronization to avoid congestion.

Non-CDP data protection is easier to implement. You can use anything from floppy disks and tape drives, all the way up to top-of-the-line public cloud services. This flexibility dramatically lowers costs, but it places a significant burden on the systems administrators to make sure the RPOs they select for their data are right.

Choosing a Data Protection Approach

As you can tell, data protection is extremely complicated. The scope of it is, quite literally, the entire scope of your organization. Not only in place, but in time, as well. Past, present and future matter just as much as location.

Any vendor who tries to sell you on the idea that one approach fits all is a vendor from which you run—not walk—away. Even the smallest of businesses will combine a Dropbox-like CDP data protection capability with some form of non-CDP protection for selected data types, even if that non-CDP protection is just periodically copying files onto a USB stick.

The key to anything in data protection is needs assessment. Work with experienced professionals to determine what your data protection needs *actually are*. Then work with vendors to see if they can meet some, or all, of those needs.

Do not shy away from putting multiple vendors into play to meet the totality of your needs. Very, very few vendors have the software, professional services expertise, hosting and public cloud services to pull off complete data protection and workload protection offerings.

Remember the importance of outside opinions. If someone's teenaged offspring—or better yet, a trained professional—can come up with hypothetical ways to poke holes in either your information security or your data protection design, no amount of justification or butt-covering paperwork will do. The problems must be solved, no matter how much work is involved.

Above all, data protection should be a fundamental part of your organization's approach to IT. It should not be an add-on or an afterthought, and under no circumstances should it be designed by a committee. Thanks to ransomware, data protection now has as much to do with information security as it does with business continuity; thus, it is the single most important element of your organization's IT.

Don't screw it up. **VR**

Trevor Pott is a full-time nerd from Edmonton, Alberta, Canada. He splits his time between systems administration, technology writing and consulting. As a consultant he helps Silicon Valley start-ups better understand systems administrators and how to sell to them.